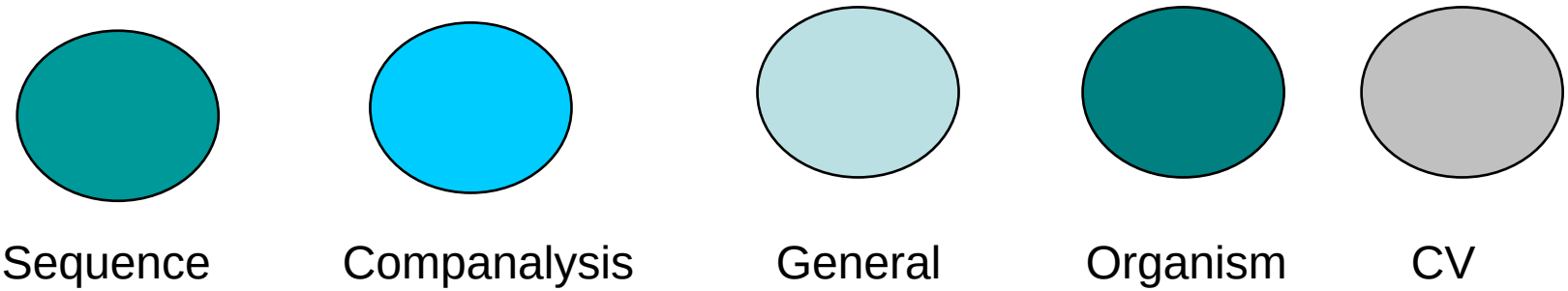
Sybil

Chado: Open Source Database

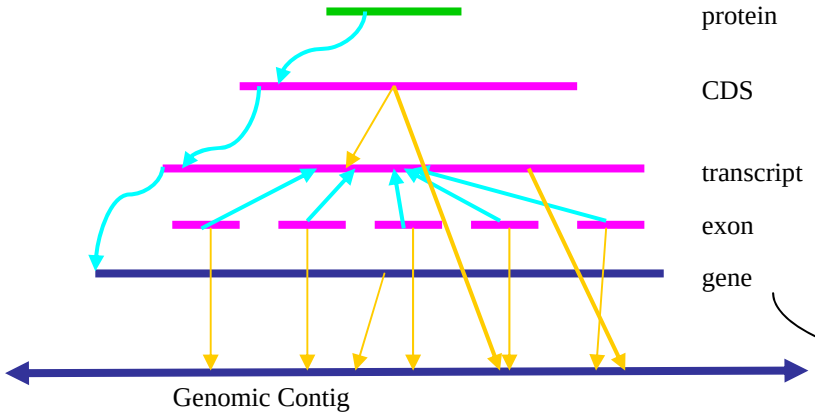


- Collaborative relational database ~15 people.
  - Flybase/Harvard
  - GMOD Consortium
- Composed of several modules
- Freely available as open source
- Many support tools under development by several laboratories. (see www.gmod.org)

Annotation attributes

Annot. Attribute terms

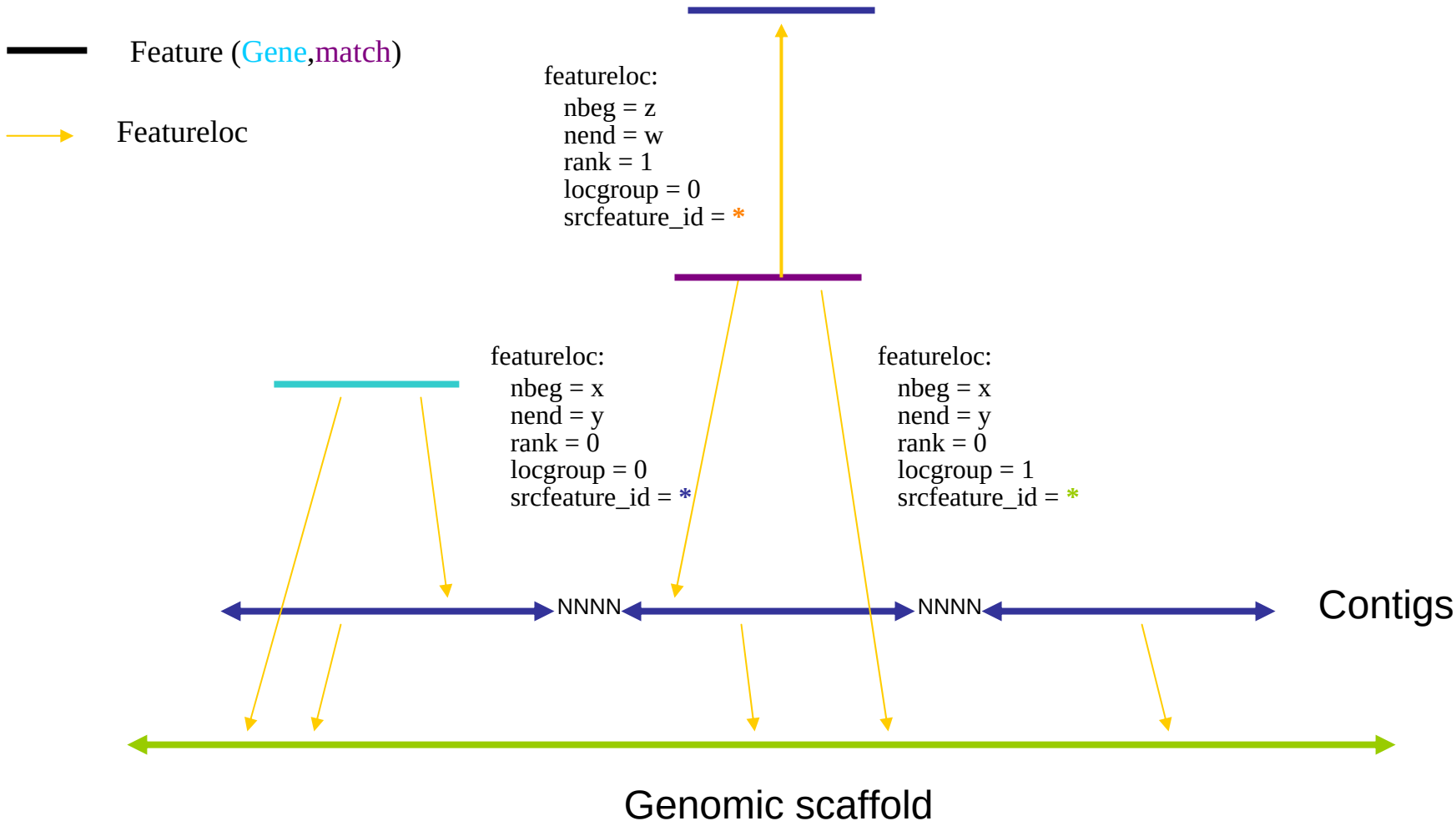
- Gene name
- EC#
- Gene sym
- GO
- …



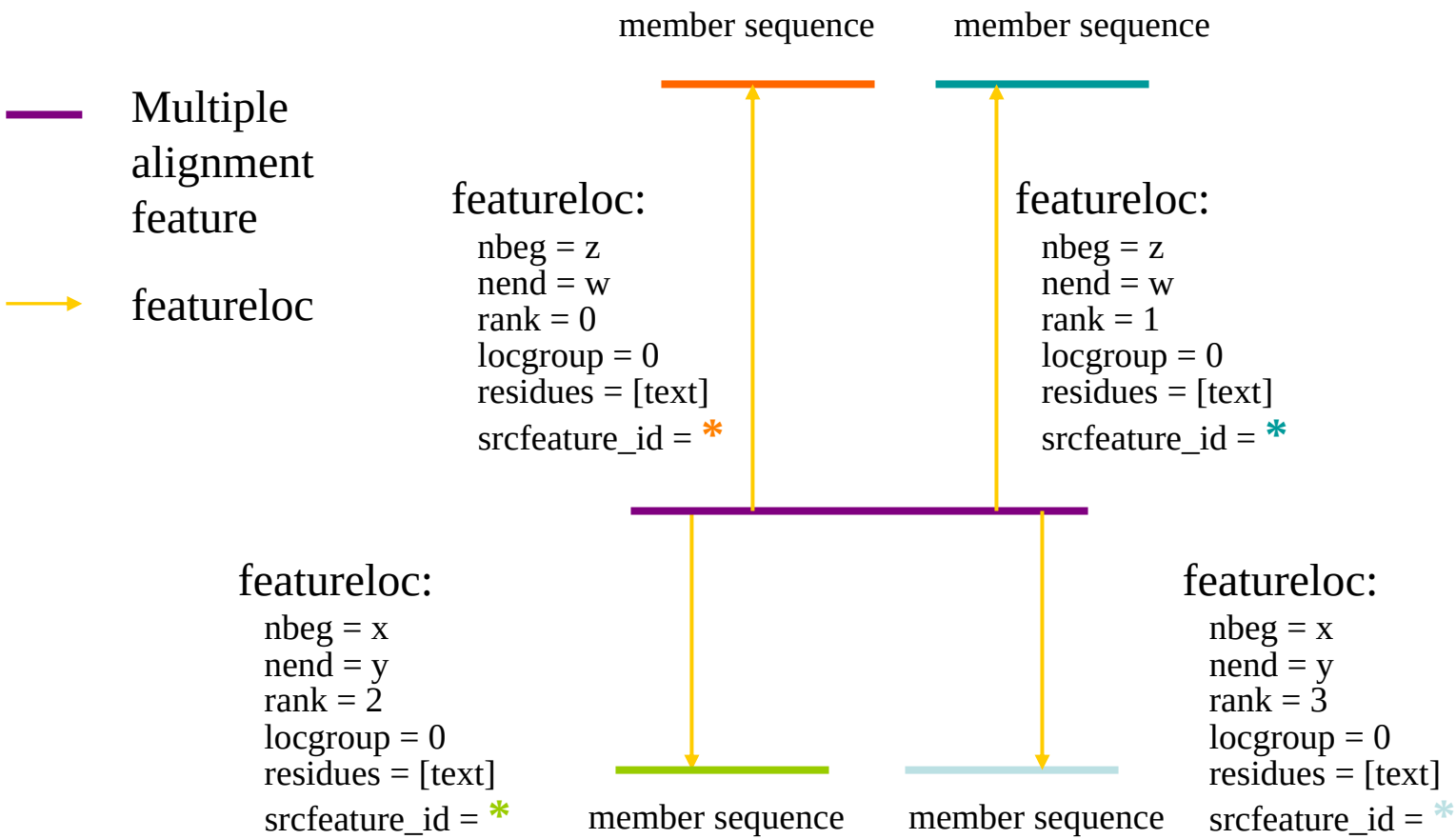
featureprop:
type_id =
value =

cvterm:
cv_id
cvterm_id
name

Scaffold mappings

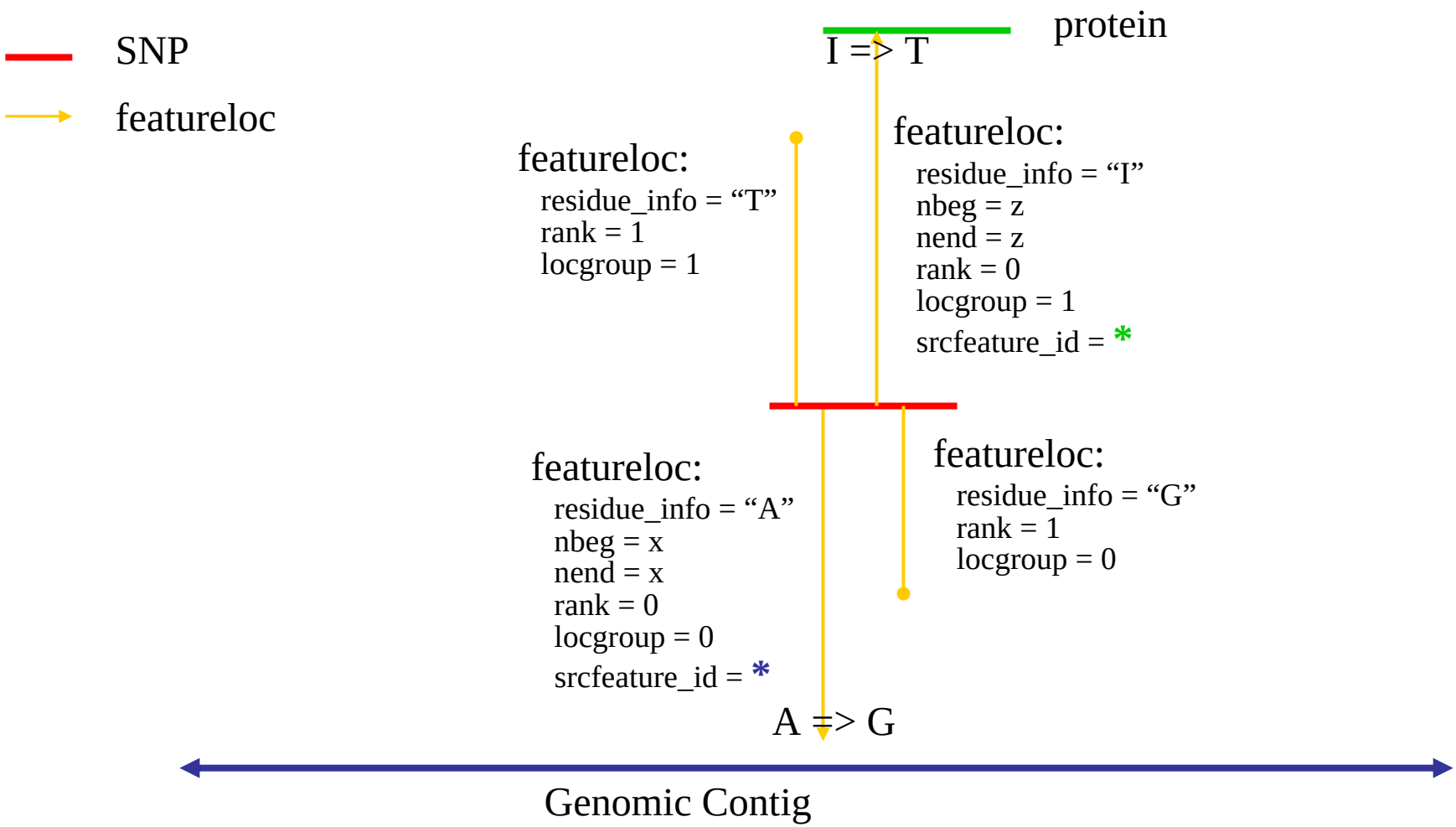


Multiple alignments



Sequence variations

SNPs (redundant mapping to protein)

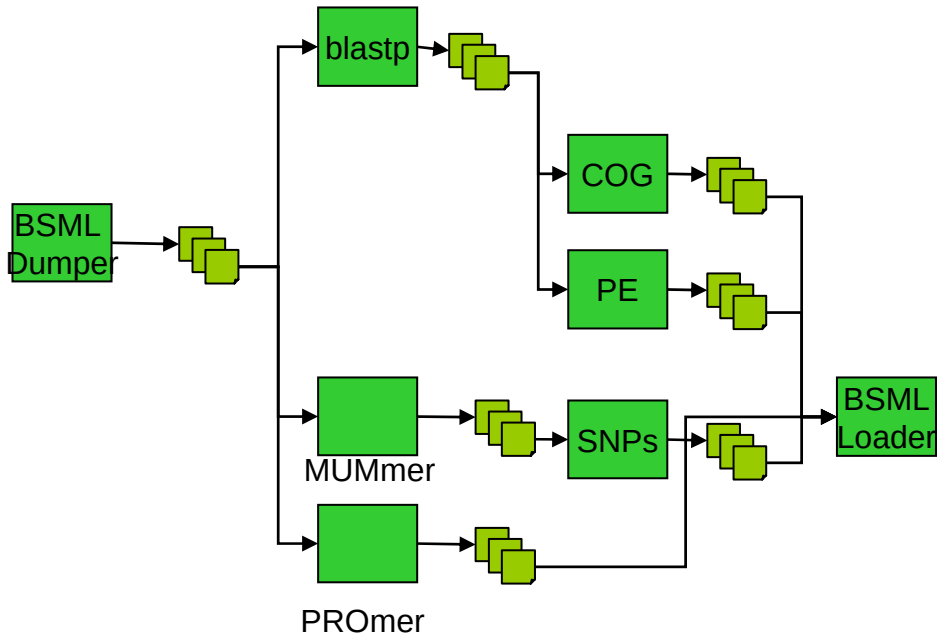


Chado

- No serious performance issues
- Many labs
- Modular
  - Sequence module is stable
- Extensive use of CV-terms
- Usage of datatypes documented

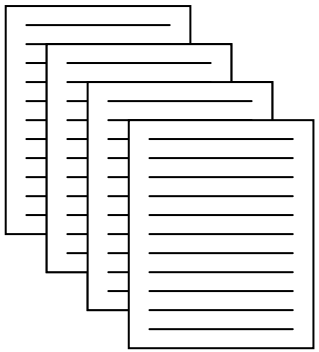
Workflow Computes

- Blast
- Position effect (conserved gene order)
- MUMmer
  - SNPs
- PROmer
- Gene families
  - COGs
  - Paralogs

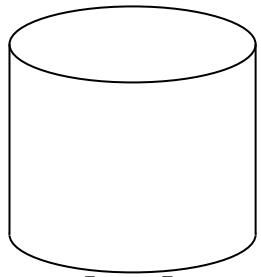


Primary output: BSML-XML

Data Prep For Comparative Analysis



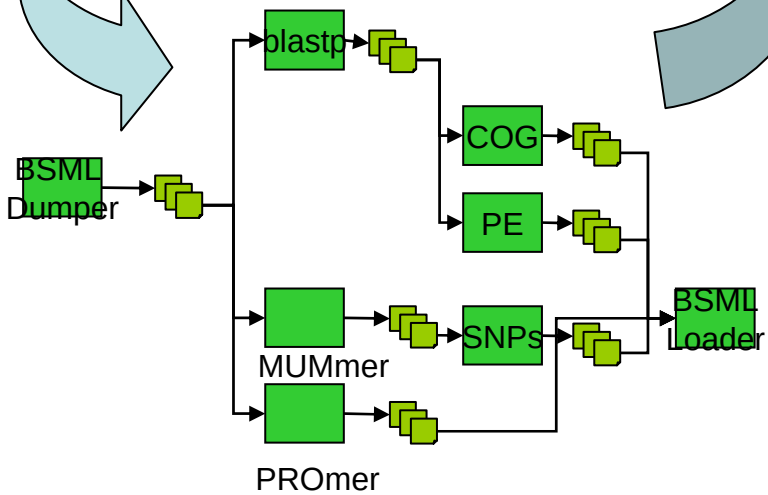
GMOD Consortium



Chado

GenBank Files
EMBL Files
Custom Files
XML
- -GFF
- -BSML

TIGR Original DB



Sybil

- **Sibyl** ancient Greek word *sibylla*, meaning prophet, someone who speaks by divine inspiration;
- Workflow system + chado db + interface

Sybil

- Workflow system → chado db
- Interface
  - Three-tiered perl architecture
  - Pushes pages to web browser
  - Many reports
  - All graphics are in svg

Sybil-searches

- All v all blastp searches
- Mummer
  - Nucleotide
  - Protein

Sybil-clustering

- Jaccard -- paralogs
  - pairs of proteins form networks
  - Edges of networks are evaluated
  - Scoring system set by user
- COGs – bidirectional best hits
- Jaccard COG-clustering
  - transitive closure
  - forms Jaccard reps from all genomes
  - multiple Jaccard reps from same genome.

Sybil-smoothing

- Position effect
  - pairwise protein matches to identify distinct sequence segments in which the gene order and orientation are conserved.
  - window_size, gap_penalty, gene_count_cutoff, gene_length_cutoff, min_matches_per_window , orientation
- Collection of genes into blocks based on J-COGs

44 BLOCK(S) OF SIZE 4

[RGAPS 0->11;1->15;2->15;3->2;4->1] [TGAPS 0->15;1->15;2->10;3->3;4->1]

1 /44	ref=bma1_14972_assembly [2321008-2362910]	tgt=cba1_35_assembly [1353380-1370854]	35.23 top
	m07286: hypothetical protein	m00530: CBG06093	
	m07287: Ras protein let-60, putative	m00535: CBG06092	
	m07292: Ankyrin repeat	m00158: CBG06095	
2 /44	ref=bma1_14972_assembly [2551960-2568777]	tgt=cba1_62_assembly [1223804-1232658]	62.5 top
	m07318: hypothetical protein	m00181: CBG17506	
	m07319: hypothetical protein	m00437: CBG17507	
	m07320: hypothetical protein	m00173: CBG17508	
	m07321: hypothetical protein	m00246: CBG17509	
3 /44	ref=bma1_14972_assembly [4167550-4179032]	tgt=cba1_87_assembly [3730641-3749526]	87.7 top
	m07534: cAMP-dependent protein kinase regulatory chain, putative	m01110: CBG14792	
	m07535: cAMP-dependent protein kinase regulatory chain, putative	m00451: CBG14791	
	m07536: Conserved hypothetical protein, putative	m01020: CBG14789	
	m07537: Fibronectin type III domain containing protein	m00132: CBG14788	
4 /44	ref=bma1_14972_assembly [4277328-4298711]	tgt=cba1_35_assembly [1088555-1106368]	35.15 top
	m07557: Conserved hypothetical protein, putative	m00139: CBG06019	
	m07558: Phosphate carrier protein, mitochondrial precursor, putative	m00269: CBG06020	
	m07560: nuclear receptor NHR-67, putative	m00081: CBG06024	
5 /44	ref=bma1_14972_assembly [4312629-4322507]	tgt=cba1_69_assembly [609782-631716]	69.7 top
	m07562: hydrolase, carbon-nitrogen family protein	m00007: CBG13313	
	m07562: hydrolase, carbon-nitrogen family protein	m00148: CBG13318	
	m07562: hydrolase, carbon-nitrogen family protein	m00112: CBG13323	
	m07563: Translocon-associated protein, delta subunit precursor, putative	m00145: CBG13316	
	m07563: Translocon-associated protein, delta subunit precursor, putative	m00050: CBG13321	
	m07564: LD29159p-related	m00061: CBG13317	
	m07564: LD29159p-related	m00052: CBG13322	
	m07565: Zinc finger, C2H2 type family protein	m00161: CBG13314	
	m07565: Zinc finger, C2H2 type family protein	m00030: CBG13319	
6 /44	ref=bma1_14992_assembly [813900-851348]	tgt=cba1_138_assembly [233661-259646]	138.3 top
	m10940: hypothetical protein	m00115: CBG11621	
	m10941: SH2 domain containing protein	m00137: CBG11620	

bma1-vs-cba1-2-4

SYNTENIC BLOCK SIZE HISTOGRAM

read 655 syntenic block(s)
read 1383 distinct syntenic reference genes
read 1381 distinct syntenic target genes
gene/matching gene counts differ for 300 block(s)
ref-tgt counts differ for 280 block(s)

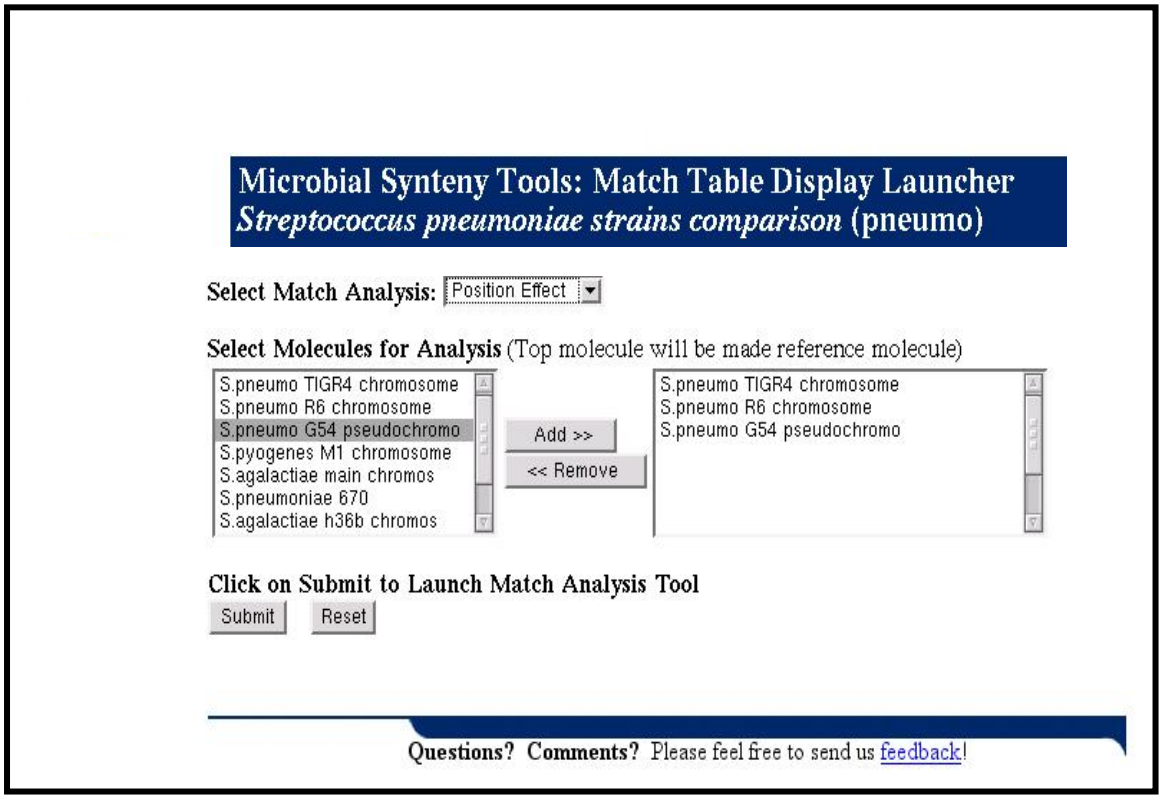
- 1 block(s) of size 8
  [rgaps 5->1] [tgaps 5->1]
- 1 block(s) of size 7
  [rgaps 2->1] [tgaps 1->1]
- 6 block(s) of size 6
  [rgaps 1->2;3->3;4->1] [tgaps 1->1;3->2;4->2;5->1]
- 16 block(s) of size 5
  [rgaps 1->8;2->4;3->2;4->2] [tgaps 0->3;1->4;2->3;3->5;5->1]
- 44 block(s) of size 4
  [rgaps 0->11;1->15;2->15;3->2;4->1] [tgaps 0->15;1->15;2->10;3->3;4->1]
- 94 block(s) of size 3
  [rgaps 0->46;1->33;2->10;3->4;4->1] [tgaps 0->48;1->27;2->14;3->3;4->2]
- 493 block(s) of size 2
  [rgaps 0->424;1->49;2->10;3->5;4->5] [tgaps 0->385;1->49;2->35;3->19;4->5]

TARGET SYNTENIC BLOCKS LISTED BY CHROMOSOME/ASSEMBLY ID

Target syntenic blocks grouped by target chromosome/assembly number:

- 1.1 1.2 1.3 1.4
- 4.1 4.2 4.3
- 5.1 5.2
- 6.1
- 7.1 7.2 7.3 7.4
- 8.1
- 9.1 9.2 9.3 9.4 9.5 9.6 9.7 9.8 9.9 9.10 9.11 9.12 9.13 9.14
- 10.1 10.2 10.3
- 11.1 11.2 11.3 11.4 11.5 11.6 11.7
- 12.1 12.2 12.3 12.4 12.5 12.6 12.7 12.8 12.9 12.10 12.11 12.12 12.13 12.14 12.15
- 13.1 13.2 13.3 13.4 13.5 13.6 13.7
- 14.1
- 17.1 17.2 17.3 17.4 17.5 17.6 17.7 17.8 17.9 17.10 17.11 17.12 17.13 17.14
- 18.1 18.2
- 19.1
- 20.1 20.2 20.3

A B C



				ORFB01863	hypothetical protein
				ORFB01865	hypothetical protein
ORF02100	conserved hypothetical protein, degenerate				
ORF02102	xanthine phosphoribosyltransferase	NTORFA1660	Xanthine phosphoribosyltransferase	ORFB01867	xanthine phosphoribosyltransferase
ORF02103	xanthine permease	NTORFA1661	Nucleobase:cation symporter for xanthine	ORFB01868	xanthine/uracil permease family protein
				ORFB01869	restriction endonuclease SsuRA
				ORFB01870	dpnA protein
				ORFB01871	DNA adenine methylase
ORF02104	DpnD protein	NTORFA1662	Restriction system of S. pneumoniae		
ORF02105	type II restriction endonuclease DpnI	NTORFA1663	Type II restriction enzyme DpnI (dpnC)		
ORF02107	conserved hypothetical protein	NTORFA1664	Conserved hypothetical protein	ORFB01872	uncharacterized domain 1, putative
ORF02110	galactose-1-phosphate uridylyltransferase	NTORFA1665	Galactose-1-phosphate uridylyltransferase	ORFB01873	galactose-1-phosphate uridylyltransferase
ORF02111	galactokinase	NTORFA1666	Galactokinase	ORFB01874	galactokinase
ORF02112	galactose operon repressor	NTORFA1667	GalR, member of GalR-LacI family of transcriptional regulators, binds DNA; regulator of gal operon	ORFB01875	sugar-binding transcriptional regulator, LacI family
ORF02113	alcohol dehydrogenase, zinc-containing	NTORFA1668	Alcohol dehydrogenase	ORFB01876	alcohol dehydrogenase, zinc-containing, putative
				ORFB01877	alcohol dehydrogenase, zinc-containing, putative

Displays

- Tri-tryp: comparative image for T. brucei chromosome 2
- Whole genome: T. brucei v. Leismania
- C. briggsae: 6.4Mb
- Strep: several strains
- Methods
- Documentation

Last comments

- Uses:
  - >10 faculty projects
    - Analysis
    - Annotation
    - In combination w/ Manatee
  - External web displays
  - Three jamboree-style projects
    - Tryps
    - Aspergilli
    - Elegans
- http://sybil.sourceforge.net/

Last comments

- Supported by
  - Chado open source db
  - Sourceforge version control
  - Enthusiastic engineers
  - BSML-XML
  - Workflow
  - Several alignment methods
  - Smoothing algorithms
  - Multiple platforms
  - API

bing badda boom.